교육과정평가연구
The Journal of Curriculum and Evaluation
2026, Vol. 29, No. 2, pp.165~190
DOI: https://doi.org/10.29221/jce.2026.29.2.165

AI 평가와 교수자 평가에서 대학생이 인식하는 평가 공정성의 차이: 평가목적, AI 신뢰, 피드백의 조절효과

김해숙 (경북대학교 박사수료)*
손원숙 (경북대학교 교수)**

요약

본 연구의 목적은 평가자 유형(AI vs 교수자)에 따른 평가 공정성 인식 차이가 평가목적(진단·형성·총괄), AI 신뢰, 피드백 제공 여부에 의해 조절되는지를 탐색하는 것이다. 이를 위해 상호독립적인 세 개의 연구를 구성하였고, 연구 대상은 4년제 대학 재학생으로 연구 1은 390명, 연구 2는 267명, 연구 3은 253명을 편의표집하였다. 구체적으로 평가자 유형과 평가목적을 조작한 대학 글쓰기 수업 평가 상황에 대한 시나리오를 제공한 후 학생들의 평가 공정성 인식을 반복 측정하였다. 분석 방법으로는 반복측정 자료의 종속성과 개인 간 차이를 고려해 개인을 무선효과로 포함한 2수준 선형 혼합모형을 적용하였다. 분석 결과, 첫째, 대학생들은 교수자 평가보다는 AI 평가에 대한 공정성을 통계적으로 유의하게 높게 인식하였으나, 이러한 관계는 평가목적에 따라 차이를 보이지 않았다. 둘째, 평가자 유형에 따른 평가 공정성의 차이는 AI 신뢰에 의하여 유의하게 조절하였는데, 즉, AI 신뢰 수준이 높을수록 AI 평가에 대한 공정성 인식이 유의하게 높았다. 셋째, 피드백이 제공된 조건에서 그렇지 않은 조건보다 AI 평가자에 대한 공정성 인식이 더 높게 나타나, 피드백의 조절효과가 확인되었다. 마지막으로 AI 기반 평가의 설계 및 적용 과정에서 공정성 인식의 형성과 관련된 다양한 심리적·맥락적 요인을 종합적으로 고려할 필요가 있음을 시사하며, 이에 대한 교육적·실천적 시사점을 논의하였다.

주제어 : AI 평가, 교수자 평가, 평가 목적, 평가 공정성, AI 신뢰, 피드백

* 제1저자, khs7278@naver.com
** 교신저자, wsohn@knu.ac.kr

I. 서 론

인공지능(Artificial Intelligence: AI)은 교육 분야를 포함한 다양한 산업의 변화를 이끌며 지난 몇 년 동안 빠르게 국내·외 환경이 변화하는데 큰 영향을 미쳤다(Chinta et al., 2024). 교육 분야에서 1950년 영국의 수학자 앨런 튜링(Alan Turing)은 “Computing Machinery and Intelligence”에서 기계의 사고 가능성을 논의하였고, 이후 John McCarthy가 제시한 ‘인공지능’이라는 용어는 1956년 다트머스 워크숍을 계기로 명명되었다. 교육분야의 AI는 1960년대와 1970년대의 컴퓨터 지원 교육에 대한 초기 실험에서 출발하여, 학습자의 개별 학습 속도와 성향을 반영한 개별화 학습을 탐구하는 방향으로 발전해 왔다(Chinta et al., 2024). 구체적으로 AI는 평가 채점의 자동화, 언어의 번역 서비스, 개인화된 학습 피드백을 제공 등 다양한 기능을 통해 학습자의 교육적 요구를 효과적으로 충족시킴으로써 교육의 질 향상과 더 나은 교육 성과를 가져올 수 있는 혁신적인 교육 기술로서 중요한 역할을 수행하고 있다(Liu et al., 2023).

한편 국내에서는 2022년 교육부가 발표한 「디지털 인재 양성 종합방안(교육부, 2022)에서 따르면 AI 기반 맞춤형 교육 플랫폼 구축, 교사의 디지털 활용 역량 강화를 중심으로 교육 분야에서의 AI 활용 방안이 추진되고 있다. 이어 2023년 디지털 기반 교육혁신 방안을 발표하며, 2028년까지 지능형 튜터링 시스템 기반의 디지털 교과서(AIDT)를 전면 도입할 계획을 제시하였다(교육부, 2023). 이후 2024년과 2025년 업무계획에서도 AI·디지털 교육전환을 위한 맞춤형 평가 체제 구축과 AI 기반 학습 분석이 지속적으로 추진되고 있음을 확인할 수 있다(교육부, 2024; 교육부, 2025). 이러한 정책 방향은 지능형 튜터링 시스템을 통해 학습자의 학습 데이터를 기반으로 피드백을 제공하고, 학습 유형 및 특성에 따른 개별화된 교수–학습 모형을 설계·적용하는 방향으로 구체화되고 있다(함은혜 외, 2024).

한편 국내 교육평가 분야에서도 AI 기반 평가 도입 사례가 점차 확대되고 있다. 대표적으로 경기도교육청의 ‘하이러닝’(경기도교육청, 2024; 김주현, 2025)은 학생의 학습 수준과 속도에 맞춘 서·논술형 자동 채점 시스템으로, 학생 답안을 분석하여 점수와 피드백을 제공함으로써 교사의 채점 부담을 경감하고 학습 지원을 강화하는 데 활용되고 있다. 또한 생성형 AI를 활용한 자동채점 시스템은 서·논술형 평가에서 채점의 일관성과 효율성을 높이는 동시에, 피드백 제공 가능성을 확장하는 방향으로 개발·적용되고 있다(정윤하, 2025). 나아가 AI 기반 자동평가 프로그램(고경민 외, 2025)을 활용하여 학생의 글쓰기 답안을 채점하고 개인별 피드백을 제공함으로써 학습자의 글쓰기 능력 향상을 지원하는 사례도 보고되고 있다.

이와 같이 AI 기술의 발전과 함께 교육 현장에서 AI 기반 평가의 활용이 확대되고 있으나, 평가의 효과성과 수용성은 기술적 정확성뿐 아니라 학습자가 지각하는 공정성에 의해 크게 좌우된다는 점에 주목할 필요가 있다(오의경, 2025; Chai et al., 2024; Venkatesh et al., 2003). 실제로 학생들이 AI 평가를 얼마나 공정하고 신뢰할 수 있는 방식으로 인식하는지는 평가 결과의 수용 수준과 교육적 효과를 결정하는 핵심 요인으로 간주된다. 특히 AI 평가는 인간 평가와 달리 판단 과정의 불투명성과 맥락 이해의 한계로 인해 공정성 인식과 관련된 새로운 쟁점을 야기할 수 있으며, AI에 대한 신뢰나 피드백 경험과 같은 심리적·맥락적 요인에 따라 동일한 평가 상황에서도 공정성 인식이 달라질 수 있다. 그럼에도 불구하고 기존 연구(박소영 외, 2023; 함은혜 외, 2024; Bui & Barrot, 2025)는 자동화 채점의 정확도나 알고리즘 비교 등 기술적 성능 검증에 주로 초점을 두어, 학생 관점에서의 공정성 인식과 평가 수용에 관한 연구는 상대적으로 부족한 실정이다. 이에 본 연구는 AI 기반 평가의 정당성과 교육적 활용 가능성을 확보하기 위해 평가 공정성 인식과 그 영향 요인을 체계적으로 규명하고자 하였다.

구체적으로 본 연구에서는 대학 글쓰기 수업 평가 상황을 기반으로 평가자 유형(AI vs 교수자)에 따른 공정성 인식의 차이를 분석하고, 평가목적, AI 신뢰, 피드백 제공 여부의 조절효과를 검증하고자 한다. 평가목적은 평가의 성격과 기능을 규정하고, AI 신뢰는 평가 결과의 수용에 영향을 미치며, 피드백 제공 여부는 평가 경험의 질과 공정성 인식 형성에 중요한 역할을 한다는 점에서 조절 변인으로 설정하였다. 이를 위해 실험적 비네트 방법(Aguinis & Bradley, 2014; Atzmüller & Steiner, 2010)을 활용하여 가상 시나리오를 제시하고 설문을 통해 공정성 인식을 측정하였으며, 반복측정 자료의 종속성과 개인 간 차이를 고려한 2수준 무선절편 선형혼합모형(Linear Mixed-effects Model with random intercepts)을 적용하여 조건별 차이와 영향 요인을 분석하였다. 한편 대학생은 평가에 대한 이해와 공정성 판단이 비교적 안정적이고, 생성형 AI 활용 경험이 풍부하여 AI 평가 상황에 대한 현실적 인식을 제공할 수 있으며, 고등교육 맥락의 다양한 평가 형태를 반영할 수 있다는 점에서 연구대상으로 선정되었다. 또한 비네트 기반 설계에 대한 이해와 응답 신뢰성 측면에서도 연구 수행에 적합한 집단으로 판단되었다.

본 연구의 구체적인 연구문제는 다음과 같다.

연구문제 1. 평가목적(진단·형성·총괄평가)은 평가자 유형(AI 알고리즘 vs 교수자)과 공정성 인식 간의 관계를 조절하는가?

연구문제 2. AI 신뢰 수준은 평가자 유형(AI 알고리즘 vs 교수자)과 공정성 인식 간의 관계를 조절하는가?

연구문제 3. 피드백 제공 여부는 평가자 유형(AI 알고리즘 vs 교수자)과 공정성 인식 간의 관계를 조절하는가?

Ⅱ. 이론적 배경

1. 평가 공정성의 개념

평가 공정성은 학습자가 평가의 과정과 결과를 얼마나 정당하고 공평하게 인식하는지를 의미하며, 이는 평가 결과 수용과 학습 참여에 중요한 영향을 미친다. 공정성은 단일 차원이 아닌 다차원적 개념으로, Colquitt(2001)는 이를 분배적·절차적·대인적·정보적 공정성으로 구분하였다. 이 중 분배적 공정성은 결과의 공정성에 대한 인식을, 대인적·정보적 공정성은 상호작용과 정보 제공의 적절성과 관련된다. 반면 절차적 공정성은 평가 과정의 공정성에 초점을 두며, 평가 기준의 일관성, 절차의 객관성, 편향 없는 적용 등에 대한 인식을 포함한다. 교육 맥락에서는 평가 결과보다 평가 과정에 대한 신뢰가 학습자의 수용성과 참여에 더 직접적인 영향을 미친다는 점에서 절차적 공정성이 중요하게 강조된다. Tierney(2012) 역시 평가 기준의 명확성과 일관성, 편견 없는 절차가 공정성 인식에 핵심적임을 지적하였다. 이는 평가 공정성이 결과를 넘어 과정 전반에 대한 인식으로 이해될 필요가 있음을 시사한다.

최근 AI 기반 평가의 확산으로 알고리즘 편향과 불투명성 등의 문제가 제기되면서 공정성 논의는 더욱 중요해지고 있다(Holmes et al., 2022). 이에 본 연구는 평가 공정성의 하위 차원 중 절차적 공정성에 초점을 두어, 평가 기준의 일관성, 절차의 투명성, 편향 여부에 대한 학습자의 인식을 중심으로 공정성을 개념화하고자 한다.

2. 평가자 유형에 따른 공정성 인식

AI 기반 평가 시스템은 학습자 데이터를 분석하여 학습 진단과 평가 정보를 실시간으로 제공하고, 방대한 데이터에서 패턴을 도출하여 학습 진행 상황을 예측함으로써 평가의 효율성과 일관성을 제고하는 데 기여한다(Luckin et al., 2016). 특히 게임 기반 활동이나 협업 프로젝트 등 다양한 학습 맥락에서 지속적인 데이터 수집과 분석을 통해 기존의 일회성 평가를 넘어선 새로운 평가 패러다임을 제시하고 있다. 이러한 흐름 속에서 AI 기반 평가 시행, 자동채점 및 피드백 시스템은 대량의 데이터를 기반으로 객관적 기준에 따라 평가를 수행하고, 즉각적인 피드백을 제공할 수 있다는 점에서 평가의 효율성과 일관성을 높이는 대안으로 주목받고 있다(김은영, 2024; 정유남, 2024; 하민수, 신철균, 2024). 그러나 이러한 기술적 장점에도 불구하고, AI 평가에 대한 공정성 인식은 여전히 중요한 쟁점으로 남아 있다. 일부 연구에서는 GPT-4 기반 자동채점 시스템의 타

당도와 일관성이 일정 수준 확보될 수 있음을 보고하고 있으나(함은혜 외, 2024), 동시에 훈련 데이터의 편향, 정답 생성의 신뢰성, 사용자 의도에 대한 맥락적 이해의 한계 등이 지적되고 있다(김은영, 2024). 특히 판단 근거가 명확하게 드러나지 않는 '블랙박스' 특성은 평가 과정의 투명성과 설명 가능성을 저해하며, 이는 학습자가 해당 평가를 얼마나 공정하고 신뢰할 수 있는 것으로 인식하는지에 직접적인 영향을 미칠 것으로 가정된다.

반면, 교수자에 의한 평가는 단순한 결과 산출을 넘어 학습자의 수준, 학습 과정, 맥락적 요인을 종합적으로 고려한 판단이 이루어진다는 점에서 특징을 갖는다. 교수자는 학습자의 이해 수준을 진단하고, 학습 과정에서 나타나는 오류나 사고 과정을 반영하여 맞춤형 피드백을 제공함으로써 평가의 교육적 기능을 강화할 수 있다. 특히 개념적 이해가 중요한 내용영역에서는 이러한 맥락 기반 피드백이 학습 효과에 중요한 영향을 미친다. 더 나아가 교수자는 교실 상호작용과 학습자의 반응을 토대로 평가 판단을 조정할 수 있으며, 이는 평가의 유연성과 맥락적 타당성을 높이는 요소로 작용한다(안혜정, 2024).

이와 같이 AI 평가와 교수자 평가는 평가 수행 방식과 판단 근거에서 구조적인 차이를 보인다. AI 평가는 데이터 기반의 일관성과 객관성을 강점으로 하는 반면, 교수자 평가는 맥락적 이해와 교육적 전문성을 바탕으로 한 해석과 판단을 특징으로 한다. 이러한 차이는 단순히 평가의 정확성 차원을 넘어, 학습자가 해당 평가를 얼마나 공정하고 정당한 것으로 인식하는지에 중요한 영향을 미칠 수 있다. 따라서 동일한 평가 결과라 하더라도 평가자 유형에 따라 학습자의 공정성 인식은 달라질 수 있으며(Chai et al., 2024), 이는 평가 결과의 수용과 활용에도 영향을 미치는 핵심 요인으로 작용할 것으로 판단된다. 기존 연구들이 주로 AI 평가의 기술적 타당성과 신뢰성 검증에 초점을 두었다면, 이제는 평가자 유형이라는 맥락적 요인이 학습자의 평가 공정성 인식에 어떠한 차이를 만들어내는지를 탐색할 필요성이 제안된다. 특히 평가자 유형에 따른 공정성 인식의 차이를 규명하는 것은 AI 기반 평가의 교육적 수용성과 활용 가능성을 이해하는 데 중요한 시사점을 제공할 것이다.

3. 평가 공정성 인식의 맥락적·심리적 조절 기제

본 연구는 학습자의 평가 공정성 인식이 다양한 맥락적·심리적 요인에 의해 조절된다는 관점을 바탕으로 한다. 특히 평가목적, AI 신뢰 및 피드백 등은 학습자가 평가 과정과 결과를 해석하고 수용하는 방식에 영향을 미친다는 점에서, 평가자 유형에 따른 공정성 인식의 차이를 설명하는 주요 조절 변인으로 선정하였다. 첫째, 평가목적은

Bloom(1969)의 분류에 따라 진단·형성·총괄 평가로 구분하였으며, 먼저 진단평가(Diagnostic Evaluation)는 학습을 시작하기 전에 학생의 학습 수준과 강·약점을 파악해, 수업 설계의 기초 자료로 활용한다(박도순, 홍후조, 2010). 반면 형성평가(Formative Evaluation)는 학습 과정에서 피드백을 제공하여 학습 결손을 보완함으로써 학습 개선을 도모하는 데 초점을 두며(Bloom et al., 1971), 총괄평가(Summative Evaluation)는 학습이 종료된 시점에 학습 성과를 종합적으로 판단하고 성취 수준을 확인하고 진로 결정에 활용된다. 즉, 각 평가목적은 학습의 현재 능력 파악, 학습 과정에서 피드백 제공, 학습 결과를 확인하는 기능을 수행한다(Chai et al., 2024). 결과적으로 평가목적 차이에 따라 학습자가 평가 상황을 어떻게 이해하는지에 영향을 미치고, 평가 결과의 중요성과 절차에 대한 공정성 판단에도 영향을 끼칠 수 있어 평가 공정성 인식의 조절하는 요인으로 작용할 것으로 가정하였다. 둘째, 학습자의 AI 신뢰(오의경, 2025; Shin, 2021)는 AI 평가 상황에서 공정성 인식 형성에 영향을 미치는 중요한 학습자 요인이다. 특히, AI 신뢰 수준은 학습자가 AI 평가 결과와 피드백을 얼마나 신뢰하고 수용하는지에 영향을 미치는 중요한 조절 요인이다. AI 신뢰는 학습자가 AI가 제시하는 정보나 평가 결과가 신뢰할 만하다고 생각하고 수용하는 태도를 의미하며 실제로 예비교사를 대상으로 한 ChatGPT 기반 메타인지 자기조절 협력학습 연구 결과, AI에 대한 지각된 신뢰 수준이 사용 태도에 긍정적인 영향을 미치는 것으로 확인되었다(Dahri et al., 2024). 셋째, 설명 정보로서의 피드백은 학습자가 자신의 학습 상태를 인식하고, 학습 과정의 오류를 수정하는데 도움을 주는 중요한 학습 자원이다(Winne & Butler, 1994). 피드백 제공은 학업 성취가 향상될 뿐만 아니라 학습동기가 유발되고, 적극적으로 학습에 참여하는 등 평가의 교육적 기능이 강화된다는 선행연구가 보고되고 있다(김재욱, 손원숙, 2021; 이빛나, 손원숙, 2018; 이빛나, 박민애, 2023; 하유라, 2021; 홍소영, 2018; Sadler, 1989). 특히 판단 과정의 불투명성이 지적되는 AI 기반 평가에서는 피드백이 이러한 한계를 보완하여 평가에 대한 신뢰와 수용성을 높이는 핵심 기제로 작용할 수 있을 것으로 판단된다.

종합적으로 볼 때, 평가자 유형에 따라 평가목적, AI 신뢰, 피드백 제공 여부가 평가 과정과 결과를 해석하고 이해하는 방식에 영향을 미치는 중요한 맥락적·심리적 요인으로 작용할 것으로 예측되며, 본 연구는 이러한 요인들이 상호작용하여 평가자 유형에 따른 평가 공정성 인식의 차이를 조절하는지 체계적으로 검토하고자 한다.

Ⅲ. 연구방법

1. 연구대상

본 연구는 세 개의 연구로 구성되었으며, 각 연구에서는 조절변인 효과를 명확히 검증하기 위해 세 개의 독립적인 표본을 활용하였다. 연구대상은 4년제 대학 대학생으로, 편의표집 방법을 사용하였으며 참여자는 연구 목적과 절차에 대한 설명을 제공받은 후 설문에 참여하였다. 연구별 연구대상의 성별과 전공계열 분포는 〈표 1〉에 제시하였다. 세 연구 모두 성별 비율에서는 비교적 균형을 이루고 있으며, 전공계열에서는 사범계열의 비율(44.3~68.9%)이 가장 높게 나타났다.

<표 1> 연구별 연구대상의 성별 및 전공계열에 따른 분포(%)

구분	범주	연구 1		연구 2		연구 3	
		빈도	(%)	빈도	(%)	빈도	(%)
성별	남자	176	(45.1)	108	(40.4)	132	(52.2)
	여자	214	(54.9)	159	(59.6)	121	(47.8)
전공계열	인문계열	30	(7.7)	22	(8.2)	21	(8.3)
	사회과학계열	13	(3.3)	10	(3.7)	11	(4.3)
	상경계열	8	(2.1)	2	(.7)	8	(3.2)
	사범계열	225	(57.7)	184	(68.9)	112	(44.3)
	자연과학계열	13	(3.3)	11	(4.1)	8	(3.2)
	공학계열	30	(7.7)	20	(7.5)	27	(10.7)
	농림생명계열	4	(1.0)	2	(.7)	4	(1.6)
	의약보건계열	11	(2.8)	3	(1.1)	11	(4.3)
	예술체육계열	55	(14.1)	13	(4.9)	50	(19.8)
	그외계열	1	(.3)	-	-	1	(.4)
전체		390	100	267	100	253	100

2. 연구 설계 및 분석 방법

가. 연구 설계

본 연구에서는 대학 글쓰기 수업 평가 상황을 기반하여 평가자 유형에 따라 대학생의 평가 공정성 인식 차이를 확인하기 위해 세 가지 연구를 설계하였다. 이를 통해 평가자 유형(AI vs 교수자)에 따른 평가 공정성 인식의 차이를 검증하고, 평가목적, AI 신뢰, 피드백 제공 여부를 조절 변인으로 설정하였다. 먼저 연구 1에서는 평가자 유형과 평가목

적에 따른 평가 공정성 인식의 차이를 검증하기 위해, 평가자 유형(2수준: AI, 교수자)과 평가목적(3수준: 진단, 형성, 총괄)을 모두 집단내 요인으로 구성한 2×3 완전 반복측정 요인설계 사용하였다. 다음으로 연구 2에서는 연구1과 동일하게 평가자 유형과 평가목적 모두를 집단내 요인으로, AI 신뢰 수준을 개인차에 따른 집단간 요인으로 구성한 혼합 요인설계를 적용하였다. 마지막으로 연구 3에서는 평가자 유형과 평가목적 모두를 집단내 요인으로, 피드백(2수준: 제공, 미제공)을 집단간 요인으로 구성한 혼합 요인설계를 적용하였다. 이와 같이 연구를 단계적으로 분리한 이유는 여러 조절변인을 단일 모형에 투입할 경우 상호작용 구조가 복잡해져 해석의 명확성이 저하되므로, 성격이 다른 조절변인을 분리함으로써 각 조절변인의 고유한 영향력을 독립적으로 검증하고 결과 해석의 명확성과 내적 타당도를 확보하기 위함이다. 또한 수업 평가 상황에서 평가목적은 핵심 변인으로 작용하므로, 세 연구의 비교와 일관성의 확보를 위해 공통적으로 집단내 요인으로 포함하였다.

나. 분석 방법

본 연구에서는 2수준 무선절편 선형혼합모형(Linear Mixed-effects Model(LMM) with random intercepts)을 적용하였는데, 이는 여러 관측값이 수집된 경우, 관측값 간의 종속성을 고려하면서 개인 간 차이를 동시에 분석할 수 있는 방법이다. 개인을 2수준, 반복측정 조건을 1수준으로 설정하고, 개인마다 서로 다른 초기 수준(절편)을 갖도록 무선효과를 포함함으로써 개인 간 이질성을 반영할 수 있다. 또한 고정효과로는 평가자 유형과 평가목적과 같은 반복측정 조건(집단내 변인)뿐 아니라 AI 신뢰나 피드백 제공 여부와 같은 개인 특성(집단간 변인)을 함께 포함할 수 있어, 동일 개인 내 조건 변화에 따른 효과와 개인 간 특성에 따른 평균 차이, 그리고 이들 간 상호작용 효과를 동시에 추정할 수 있다는 장점이 있다.

이에 따라 각 연구문제에 적용한 모형은 다음과 같다. 연구 1에서는 평가자 유형과 평가목적의 효과를 검증하기 위하여 다음의 모형을 설정하였다.

$$Y_{ij} = \gamma_{00} + \gamma_{10} AI_{ij} + \gamma_{20}\text{진단}_{ij} + \gamma_{30}\text{형성}_{ij} + \gamma_{40}(AI \times \text{진단})_{ij} + \gamma_{50}(AI \times \text{형성})_{ij} + u_{0j} + r_{ij}$$

연구 2에서는 연구 1의 기본 모형에 AI 신뢰 변인을 추가하여 다음과 같이 분석하였다.

$$Y_{ij} = \gamma_{00} + \gamma_{10} AI_{ij} + \gamma_{20}\text{진단}_{ij} + \gamma_{30}\text{형성}_{ij} + \gamma_{40}(AI \times \text{진단})_{ij} + \gamma_{50}(AI \times \text{형성})_{ij} + \gamma_{01}\text{신뢰}_{j} + \gamma_{11}(AI_{ij} \times \text{신뢰}_{j}) + {}_{21}(\text{진단}\gamma_{ij} \times \text{신뢰}_{j}) + \gamma_{31}(\text{형성}_{ij} \times \text{신뢰}_{j}) + \gamma_{41}(AI_{ij} \times \text{진단}_{ij} \times \text{신뢰}_{j}) + \gamma_{51}(AI_{ij} \times \text{형성}_{ij} \times \text{신뢰}_{j}) + u_{0j} + r_{ij}$$

연구 3에서는 연구 1의 기본 모형에 피드백 제공 여부 변인을 추가하여 다음과 같이 분석하였다.

$$Y_{ij} = \gamma_{00} + \gamma_{10} AI_{ij} + \gamma_{20}\text{진단}_{ij} + \gamma_{30}\text{형성}_{ij} + \gamma_{40}(AI \times \text{진단})_{ij} + \gamma_{50}(AI \times \text{형성})_{ij} + \gamma_{01}\text{피드백}_{j} + \gamma_{11}(AI_{ij} \times \text{피드백}_{j}) + \gamma_{21}(\text{진단}_{ij} \times \text{피드백}_{j}) + \gamma_{31}(\text{형성}_{ij} \times \text{피드백}_{j}) + \gamma_{41}(AI_{ij} \times \text{진단}_{ij} \times \text{피드백}_{j}) + \gamma_{51}(AI_{ij} \times \text{형성}_{ij} \times \text{피드백}_{j}) + u_{0j} + r_{ij}$$

여기서 Y_{ij}는 개인 j의 시나리오 i에 대한 평가 공정성 인식 점수, u_{0j}는 개인 간 차이를 반영하는 무선절편, r_{ij}는 수준 1의 오차항을 의미한다. 평가목적에서는 총괄평가=0, 평가자 유형에서는 교수자=0으로 코딩하였다. 구체적으로 연속변인인 AI 신뢰의 경우 평균 중심화(mean centering)하였고, 범주 갯수(k)가 3개 이상인 범주형 변인은 $(k-1)$개의 더미 변인으로 변환하였다. 한편 상호작용 효과가 유의한 경우, 사후분석으로 상호작용 효과 그래프 및 단순효과(simple effect) 분석을 실시하였다. 본 연구의 분석은 IBM SPSS Statistics 25.0를 사용하여 수행하였다.

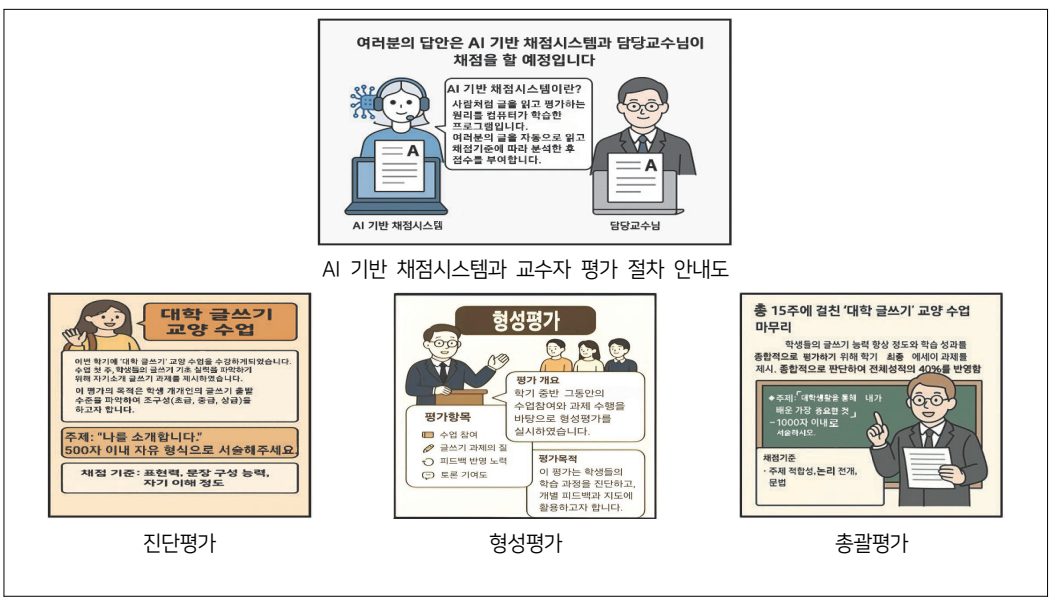
3. 측정 도구

가. 평가 상황 시나리오

본 연구에서는 대학생의 평가 공정성 인식을 측정하기 위해 실험적 비네트 방법론(Experimental Vignette Methodology, EVM)을 활용하였다. EVM은 연구 목적에 따라 설계된 현실적인 상황 시나리오를 제시하고, 이에 대한 참여자의 인식이나 판단을 측정하는 방법으로, 변인을 체계적으로 조작함으로써 인과적 해석이 가능하다는 장점이 있다(Aguinis & Bradley, 2014). 이와 같이 시나리오 기반 척도(Ekici et al., 2023)는 구체적 상황을 제시함으로써 추상적 자기보고식 응답의 한계를 보완하고, 응답자가 상황을 해석하는 인지적 과정을 거치도록 하여 문항 이해와 응답의 타당성을 높인다. 또한 상황 판단을 선행하게 함으로써 응답의 방향성을 직관적으로 파악하기 어렵게 만들어 사회적 바람직성에 따른 응답 편향을 완화하는 것으로 보고되고 있다(권은아, 이종구, 2020; 백순근 외, 2024).

본 연구는 실제로 경험할 수 있는 대학 글쓰기 수업 평가 상황을 반영한 내용을 기반으로 평가목적(진단·형성·총괄)에 따라 3개의 평가 상황 시나리오를 구성하였다. 예를 들어, 진단평가 시나리오에서는 학기 초 맞춤형 학습 지원을 위한 글쓰기 기초 능력을 평가하기 위한 상황, 형성평가 시나리오는 학기 중 수업 참여도, 과제에 대한 피드백, 토론 기여도 등을 반영하여 학습 과정을 진단하는 상황, 총괄평가 시나리오는 학기 말에

글쓰기 역량을 종합적으로 평가하기 위해 서술형 글쓰기 과제로 학습 성취 수준을 평가하고, 그 결과를 최종 성적에 반영하는 상황을 제시하였다([그림 1]). 한편 평가 정보를 평가 개요, 목적, 항목의 순서로 제시하여 참여자가 상황을 단계적으로 파악할 수 있도록 시나리오를 구성하였다.



[그림 1] 진단, 형성, 총괄평가 상황 시나리오 예시

한편 연구 3에서는 연구 1, 2의 기본 시나리오를 유지한 상태에서, 〈표 2〉와 같이 피드백 조건별로 서로 다른 평가결과가 제공되었다.

<표 2> 피드백 조건별로 제공된 평가결과

피드백 제공 조건	
제공	**미제공**
• 최종 점수: 80점 • **표현력**: 85점 (우수) • **문장 구성 능력**: 78점 (양호) • **자기 이해 정도**: 77점 (양호)	
• **표현력**: 어휘 선택이 적절하며 다양한 표현을 교과서를 참고하여 활용해 보세요. • **문장 구성 능력**: 문장 간 연결이 자연스럽지 않음. 접속어를 활용해 자연스러운 흐름을 만들어 보세요. • **자기 이해 정도**: 구체적 경험을 서술하면 자기 이해가 더욱 잘 표현되니 활용해 보세요.	• 최종 점수: 80점

나. 평가 공정성 인식

대학생들의 평가의 공정성 인식을 측정하기 위해 Colquitt(2001)가 개발한 조직공정성 척도를 활용하였다. 평가 공정성은 평가가 일관된 기준에 근거하여 적용되고, 편견이나 사적 판단 없이 객관적으로 이루어지며, 학습자의 성취와 노력을 정당하게 반영할 것이라는 신뢰 수준으로 정의하였고, 원척도의 4개 하위요인 중 절차적 공정성 요인을 측정하는 총 4개 문항을 사용하였다. 각 문항은 5점 Likert식 척도로 측정되었으며 척도 문항과 척도의 신뢰도(Cronbach's α)는 〈표 3〉과 같다. 척도 신뢰도는 평가목적 및 평가자 유형별로 제시하였고, 연구별로 산출된 신뢰도의 범위(최소값~최대값)를 제시하였다. 평가자 유형별로 보면, AI 평가는 .803～.885, 교수자 평가는 .794～.870 범위로 나타나 전반적으로 유사한 수준의 신뢰도를 보였으나, 총괄평가에서는 AI 평가자가 교수자 평가보다 다소 높은 경향을 보였다. 한편 평가목적별로는 총괄평가에서 가장 높고 진단평가에서 가장 낮은 경향이 나타났다.

<표 3> 평가 공정성 인식의 척도 문항과 신뢰도

척도 문항	신뢰도(최소 ~최대)		
	평가 목적	AI	교수자
나는 이 평가 방식이 공정하다고 느낀다.	진단	.803 ~ .805	.794 ~ .812
나는 이 평가자가 일관된 기준을 적용한다고 생각한다.	형성	.835 ~ .847	.849 ~ .853
나는 이 평가 방식은 편견 없이 객관적으로 이루어질 것이라고 생각한다.	총괄	.875 ~ .885	.816 ~ .870
나는 이 평가 방식이 나를 정당하게 평가할 것으로 생각한다.	계	.803 ~ .885	.794 ~ .870

다. AI 신뢰

AI 신뢰는 윤가율(2024)의 AI 신뢰 척도(총 4개 문항)로 측정하였고, 학습자가 AI가 제공하는 정보들을 신뢰할 만하다고 인식하는 정도로 정의하였다. 본 연구 목적에 부합하도록 일부 문구를 수정하여 사용하였는데, 예를 들어 정보의 제공 주체가 특정되지 않은 경우 가독성이 저하될 수 있으므로 "AI가"와 같은 주어를 추가하였다. 각 문항은 5점 Likert식 응답척도로 측정되었고, 신뢰도는 .718로 양호한 수준의 내적 일관성을 보였다(〈표 4〉 참고).

<표 4> AI 신뢰의 척도 문항

척도 문항
나는 AI가 제공하는 정보를 신뢰할 수 있다.
나는 AI가 제공하는 정보의 출처를 신뢰할 수 있다.
나는 AI가 신뢰할 수 있는 정보를 제공하는 능력을 믿는다.
AI는 내가 요청하는 작업을 수행할 수 있는 능력이 있다고 생각한다.

IV. 연구결과

1. 연구 1

연구 1에서는 평가자 유형(AI vs 교수자)과 평가목적(진단·형성·총괄평가)에 따라 대학생의 평가 공정성 인식의 차이를 검증하였다(〈표 5〉 참고). 먼저 무선효과 분석 결과, 개인 간 분산(τ_{00})은 .18, 잔차 분산(σ^2)은 .29로 나타났으며, ICC는 .38로 산출되었다. 이는 공정성 인식 분산의 약 38%가 개인 간 차이에 의해 설명되며, 이는 반복측정 자료의 종속성을 고려한 다층모형 적용이 적절함을 시사한다. 다음으로 고정효과 분석 결과, 평가자 유형과 평가목적 간 상호작용 효과(평가자*진단(β = .02), 평가자*형성(β = .00))는 모두 통계적으로 유의하지 않은 것으로 나타났다(p 〉 .05). 반면, 평가자 유형의 주효과(β = .17, p 〈 .05)는 통계적으로 유의하게 나타났고, Cohen's d = .31로 중간 수준의 효과크기를 보였다. 이는 AI 평가자가 교수자에 비해 더 공정하게 인식되는 경향이 존재하나, 그 차이는 비교적 완만한 수준임을 시사한다. 평가목적의 경우, 진단평가는 총괄평가에 비해 공정성 인식이 통계적으로 유의하게 낮았지만(β = −.11, p 〈 .05), 효과크기 Cohen's d = −.20 수준으로, 그 차이는 작은 수준에 해당한다. 한편 형성평가와 총괄평가 간 차이는 통계적으로 유의하지 않았다(p 〉 .05). 요약하면, 대학생들은 교수자 평가보다는 AI 평가에 대한 공정성을 보다 높게 인식하였고, 이는 평가목적에 따라 차이를 보이지 않았다.

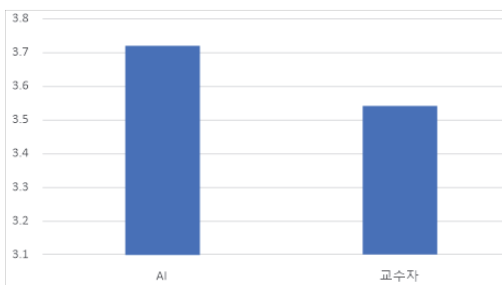
<표 5> 평가자 유형과 평가목적에 따른 공정성 인식의 차이 분석: 2수준 LMM 분석결과

고정효과	추정치(β)	SE	t	p	95% 신뢰구간(CI)
절편(0=교수자, 총괄)	3.58	.03	103.28	.00	[3.51, 3.65]
평가자 유형(1=AI)	**.17**	**.04**	**4.41**	**.00**	**[.09, .25]**
평가목적(1=진단)	**-.11**	**.04**	**-2.97**	**.00**	**[-.19, -.04]**

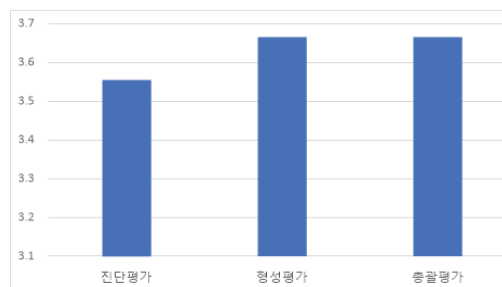
고정효과	추정치(β)	SE	t	p	95% 신뢰구간(CI)
평가목적(1=형성)	.00	.04	-.05	.96	[-.08, .07]
평가자 유형(1=AI)×평가목적(1=진단)	.02	.05	.35	.73	[-.09, .13]
평가자 유형(1=AI)×평가목적(1=형성)	.00	.05	-.04	.97	[-.11, .11]
무선효과	**분산**	**SE**			**95% 신뢰구간(CI)**
개인 무선절편 분산(τ00)	.18	.02			[.15, .21]
잔차 분산(σ2)	.29	.01			[.27, .31]
ICC	.38				

주: 참조범주(=0)는 교수자, 총괄; ICC = τ_{00} / (τ_{00} + σ^2)

한편 평가자 유형과 평가목적에 따른 평가 공정성 인식 양상은 [그림 2]과 [그림 3]에 제시하였다. 평가자 유형에서는 AI 평가의 공정성 인식이 교수자 평가 보다 높게 나타났고, 평가목적별 평가 공정성 인식은 진단평가에 비해 형성 및 총괄평가에서 높게 나타났으나, 통계적으로는 진단평가와 총괄평가 간에서만 유의한 차이가 나타났다. 이는 학습자가 평가의 목적과 결과 활용 가능성이 보다 분명한 총괄평가 상황에서 공정성을 더 높게 인식하는 경향을 시사한다.



[그림 2] 평가자 유형별 평가 공정성 인식



[그림 3] 평가목적별 평가 공정성 인식

2. 연구 2

연구 2에서는 평가자 유형별 평가 공정성 인식 양상이 평가목적 및 AI 신뢰에 따라 차이를 보이는지 검증하였다(〈표 6〉 참고). 먼저 무선효과 분석 결과, 개인 간 분산(τ_{00})은 .17, 잔차 분산(σ^2)은 .29, ICC는 .37로 연구 1과 유사하게 개인 간 이질성이 비교적 크게 나타났다. 다음으로 고정효과 분석 결과, 평가자 유형, 평가목적, AI 신뢰의 삼원 상호작용 효과(평가자*진단*AI 신뢰(β = −.08), 평가자*형성*AI 신뢰(β = .01))는 통계적으로 유의하지 않았다(p 〉 .05). 또한 평가자 유형과 평가목적 간 상호작용 효과(평가자*진단(β = −.03), 평가자*형성(β = −.02)), 평가목적과 AI 신뢰 간 상

호작용 효과(진단*AI 신뢰(β = .08), 형성*AI 신뢰(β = −.02)) 역시 모두 통계적으로 유의하지 않은 것으로 나타났다(p 〉 .05). 반면, 평가자 유형과 AI 신뢰의 상호작용 효과는 통계적으로 유의하게 나타났고(평가자*AI 신뢰(β = .47, p 〈 .05)), 효과크기 Cohen's d = .87로, 큰 수준의 효과크기를 보였다. 이는 AI 신뢰 수준에 따라 AI 평가자와 교수자 평가자 간의 공정성 인식의 차이가 비교적 뚜렷하게 달라짐을 시사한다.

<표 6> 평가자 유형, 평가목적, AI 신뢰에 따른 공정성 인식의 차이 분석: 2수준 LMM 분석결과

고정효과	추정치(β)	SE	t	p	95% 신뢰구간(CI)
절편(0=교수자, 총괄, AI 신뢰)	3.59	.06	6.71	.00	[3.47. 3.70]
평가자 유형(1=AI)	**.20**	**.07**	**3.08**	**.00**	**[.07, .33]**
평가목적(1=진단)	-.03	.07	-.49	.62	[-.16, .10]
평가목적(1=형성)	.02	.07	.23	.82	[-.11, .15]
평가자 유형(1=AI)×평가목적(1=진단)	-.03	.09	-.35	.73	[-.22, .15]
평가자 유형(1=AI)×평가목적(1=형성)	-.02	.09	-.19	.85	[-.20, .17]
AI 신뢰	-.03	.09	-.33	.74	[-.21, .15]
평가자 유형(1=AI)×AI 신뢰	**.47**	**.10**	**4.58**	**.00**	**[.27, .67]**
평가목적(1=진단)×AI 신뢰	.08	.10	.74	.46	[-.12, .28]
평가목적(1=형성)×AI 신뢰	-.02	.10	-.17	.87	[-.22, .18]
평가자 유형(1=AI)×평가목적(1=진단)×AI 신뢰	-.08	.14	-.54	.59	[-.36, .21]
평가자 유형(1=AI)×평가목적(1=형성)×AI 신뢰	.01	.14	.04	.97	[-.28, .29]
무선효과	**분산**	**SE**	**95% 신뢰구간(CI)**		
개인 무선절편 분산(τ00)	.17	.03	[.12, .23]		
잔차 분산(σ2)	.29	.02	[.26, .32]		
ICC	.37				

주: 참조범주(=0)는 교수자, 총괄; AI 신뢰는 평균중심화함; ICC = τ_{00} / (τ_{00} + σ^2)

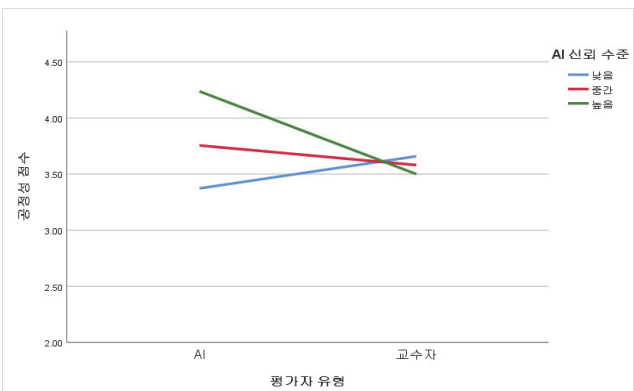
평가자 유형과 AI 신뢰의 상호작용 효과에 대한 사후분석으로 단순효과분석을 실시하였다(〈표 7〉참고). 분석 결과, AI 신뢰 수준이 낮은 집단에서는 교수자(M = 3.659)에 대한 평가 공정성을 AI (M = 3.373)보다 통계적으로 유의하게 높게 인식하는 것으로 나타났다(p 〈 .05). 반면 AI 신뢰 수준이 중간인 집단에서는 AI(M = 3.756)에 대한 평가 공정성 인식이 교수자(M = 3.581)의 경우 보다 통계적으로 유의하게 높았다(p 〈 .05). 동일하게 AI 수준이 높은 집단에서도 AI (M = 4.237)에 대한 평가 공정성 인식이 교수자(M = 3.500)보다 통계적으로 유의하게 높았고(p 〈 .05), 세 수준 중 가장 높은 평가 공정성 인식을 나타내었다.

<표 7> AI 신뢰 수준에 따른 평가자 유형의 단순효과 분석 결과(N=267)

AI 신뢰 수준	평가자 유형	평균	표준오차	$F_{(1,1331)}$	*p*
낮음	AI	3.373	0.116	8.857	.003
	교수자	**3.659**	**0.116**		
중간	**AI**	**3.756**	**0.056**	14.098	.000
	교수자	3.581	0.056		
높음	**AI**	**4.237**	**0.119**	56.204	.000
	교수자	3.500	0.119		

주: 평균은 혼합효과모형에 근거한 추정 주변평균임.

평가자 유형과 AI 신뢰 수준 간에 상호작용 양상을 살펴보기 위하여 AI 신뢰는 평균 중심화한 연속변인을 활용하였다, AI 신뢰의 평균은 3.21(SD = .65)로 나타났으며, 상호작용 효과를 시각화하기 위해 평균±1SD 기준에 따라 낮은 수준(2.56이하), 중간 수준(2.56초과~3.86미만), 높은 수준(3.86 이상)으로 구분하여 그래프를 제시하였다([그림 4] 참고). AI 평가자 조건에서는 AI 신뢰 수준이 높을수록 공정성 인식이 증가하는 경향이 나타난 반면, 교수자 평가자 조건에서는 AI 신뢰 수준에 따른 공정성 인식의 차이가 거의 나타나지 않았다. 이는 AI 신뢰가 공정성 인식에 미치는 영향이 평가자 유형에 따라 달라지며, 특히 AI 평가 상황에서 그 영향이 보다 뚜렷하게 나타남을 시사한다.



[그림 4] 평가자 유형과 AI 신뢰 수준에 따른 평가 공정성 인식의 상호작용

3. 연구 3

연구 3에서는 평가자 유형(AI vs 교수자)과 평가목적(진단·형성·총괄평가), 피드백 제공 여부에 따라 대학생의 평가 공정성 인식이 차이가 있는지 검증하였다(〈표 8〉 참고). 연구1 ,2와 동일하게 ICC는 .36로, 공정성 인식에서 개인 간 이질성이 크게 나타났다.

다음으로 고정효과 분석 결과, 평가자 유형, 평가목적, 피드백의 삼원 상호작용 효과(평가자*진단*피드백(β = .00), 평가자*형성*피드백(β = −.01)) 모두 통계적으로 유의하지 않았다(p 〉 .05). 또한 평가자 유형과 평가목적 간 상호작용 효과(평가자*진단(β = .00), 평가자*형성(β = .01)), 평가목적과 피드백 간 상호작용 효과(진단*피드백(β = −.06), 형성*피드백(β = −.06)) 역시 모두 통계적으로 유의하지 않은 것으로 나타났다(p 〉 .05). 반면 평가자 유형과 피드백 간 상호작용 효과(β = .26, p 〈 .05)는 통계적으로 유의하게 나타났고, 효과크기 Cohen's d = .48 수준으로, 중간 수준에 가까운 효과크기를 보였다. 이는 피드백 제공 여부에 따라 AI 평가자와 교수자 간 공정성 인식의 차이가 달라졌음을 시사한다. 한편 평가자 유형, 평가목적 및 피드백의 주효과는 모두 통계적으로 유의하지 않았다(p 〉 .05).(〈표 8〉 참조).

<표 8> 평가자 유형, 평가목적. 피드백에 따른 공정성 인식의 차이 분석: 2수준 LMM 분석결과

고정효과	추정치(β)	SE	t	p	95% 신뢰구간(CI)
절편(0=교수자, 총괄, 피드백무)	3.68	.06	61.87	.00	[3.56, 3.79]
평가자 유형(1=AI)	.08	.07	1.14	.26	[-.06, .21]
평가목적(1=진단)	-.03	.07	-.47	.64	[-.16, .10]
평가목적(1=형성)	.01	.07	.15	.88	[-.12, .14]
평가자 유형(1=AI)×평가목적(1=진단)	.00	.09	-.02	.98	[-.19, .18]
평가자 유형(1=AI)×평가목적(1=형성)	.01	.09	.06	.95	[-.18, .19]
피드백(1=피드백유)	-.15	.08	-1.73	.09	[-.31, .02]
평가자 유형(1=AI)×피드백(1=피드백유)	**.26**	**.10**	**2.68**	**.01**	**[.07, .44]**
평가목적(1=진단)×피드백(1=피드백유)	-.06	.10	-.62	.54	[-.25, .13]
평가목적(1=형성)×피드백(1=피드백유)	-.06	.10	-.61	.55	[-.25, .13]
평가자 유형(1=AI)×평가목적(1=진단)×피드백(1=피드백유)	.00	.14	.01	.99	[-.26, .27]
평가자 유형(1=AI)×평가목적(1=형성)×피드백(1=피드백유)	-.01	.14	-.10	.92	[-.28, .25]
무선효과	**분산**	**SE**	**95% 신뢰구간(CI)**		
개인 무선절편 분산(τ00)	.16	.02	[.13, .21]		
잔차 분산(σ2)	.29	.01	[.27, .31]		
ICC	.36				

주: 참조범주(=0)는 교수자, 총괄, 피드백무; ICC = τ_{00} / (τ_{00} + σ^2)

한편 평가자 유형과 피드백 제공 여부의 상호작용 효과에 대한 사후분석으로 단순효과분석을 실시하였다(〈표 9〉 참고). 분석 결과, 먼저 피드백 제공 조건에서는 평가자 유

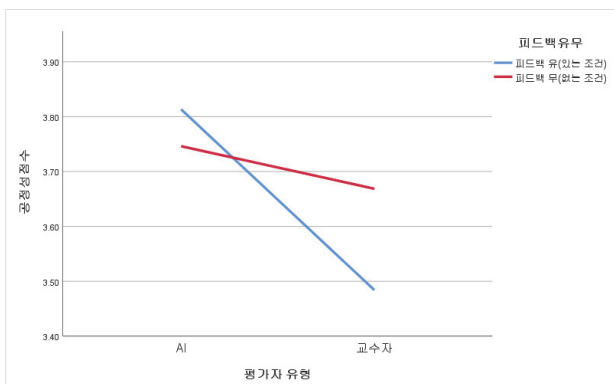
형에 따른 평가 공정성 인식의 차이가 통계적으로 유의하게 나타났다(F = 70.724, p < .05). 즉, AI 평가자에 대한 공정성 인식(M = 3.813)이 교수자 평가자(M = 3.484)의 경우 보다 높았다. 또한 피드백 미제공 조건에서도 평가자 유형에 따른 평가 공정성 인식의 차이가 통계적으로 유의하게 나타났다(F = 4.008, p < .05). 즉, AI 평가자에 대한 공정성 인식(M = 3.746)이 교수자 평가자(M = 3.669)보다 높게 나타났으나 이 차이는 피드백 제공 조건보다 작은 수준이었다.

<표 9> 피드백 제공 여부에 따른 평가자 유형의 단순효과 분석 결과(N=253)

피드백	평가자 유형	평균	표준오차(SE)	$F_{(1,1263)}$	p
제공	AI	3.813	0.046	70.724	.000
	교수자	3.484	0.046		
미제공	AI	3.746	0.045	4.008	.046
	교수자	3.669	0.045		

주: 평균은 혼합효과모형에 근거한 추정 주변평균임.

[그림 5]의 평가자 유형과 피드백 제공 여부 간의 상호작용 효과를 살펴본 결과, AI 평가자에 대한 공정성 인식은 피드백이 제공된 경우가 피드백이 제공되지 않은 경우보다 더 높은 점수를 보였다. 반면 교수자 평가 조건에서는 피드백 제공 여부에 따른 점수 차이가 감소하거나 오히려 피드백이 제공된 경우 점수가 더 낮아지는 경향이 나타났다. 이러한 결과는 피드백 제공의 효과가 평가자 유형에 따라 다르게 나타남을 시사한다.



[그림 5] 평가자 유형과 피드백 제공 여부에 따른 공정성 인식

V. 논의

본 연구는 AI와 교수자 평가 상황에서 대학생이 인식하는 평가 공정성 차이를 검증하고, 이러한 관계에서 평가목적, AI 신뢰, 피드백 제공 여부의 역할을 규명하고자 하였다. 본 연구의 결과를 논의하면 다음과 같다.

첫째, 대학교 글쓰기 평가 상황에서 AI 평가가 교수자 평가보다 공정성 인식이 전반적으로 더 높게 나타났으며, 이러한 경향은 진단·형성·총괄평가 전반에 걸쳐 일관되게 확인되었다. 이는 학습자가 평가 공정성을 판단할 때 평가자 유형이 지닌 절차적 특성, 특히 객관성과 일관성에 대한 기대를 중요한 기준으로 활용하고 있음을 시사한다. 선행연구(Chai et al., 2024)에 따르면, AI 평가는 동일한 알고리즘과 기준에 따라 평가가 이루어진다는 점에서 인간 평가에서 발생할 수 있는 주관성이나 편향을 상대적으로 배제할 수 있는 방식으로 인식되며, 이러한 특성이 공정성 인식을 높이는 요인으로 작용할 수 있음을 제안하였다. 또한 평가 기준의 일관된 적용과 신속한 결과 제공은 평가 과정의 투명성과 예측 가능성을 높여 학습자의 수용성을 강화하는 데 기여할 수 있음을 시사한다. 반면 교수자 평가는 학습자의 상황과 맥락을 반영한 판단이 가능하다는 장점이 있음에도 불구하고, 평가 기준의 적용이 상황에 따라 달라질 수 있다는 인식으로 인해 상대적으로 덜 일관된 평가로 지각될 가능성이 있을 것이다. 이러한 결과는 기술 기반 평가가 지닌 객관성에 대한 기대가 학습자의 공정성 인식에 긍정적으로 작용할 수 있음을 보여주며, 동시에 평가 공정성은 평가 과정에 대한 인식과 신뢰에 의해 형성된다는 점을 시사한다.

둘째, AI 신뢰 수준에 따라 평가자 유형에 대한 공정성 인식의 양상이 다르게 나타났다. 구체적으로 AI 신뢰가 낮은 집단에서는 교수자 평가를 AI 평가보다 더 공정하게 인식한 반면, AI 신뢰가 중간 및 높은 집단에서는 AI 평가를 교수자 평가보다 더 공정하게 인식하는 경향이 확인되었다. 특히 AI 신뢰가 높을수록 두 평가자 유형 간 공정성 인식의 차이는 더욱 뚜렷하게 나타났다. 이러한 결과는 AI 평가에 대한 공정성 인식이 단순히 평가자 유형 자체에 의해 형성되기보다는, 학습자가 AI를 얼마나 신뢰하는지와 같은 개인의 심리적 요인에 의해 크게 영향을 받을 수 있음을 시사한다. 이는 AI에 대한 신뢰 수준과 같은 개인의 태도와 신념이 기술 기반 의사결정에 대한 수용성과 평가를 좌우한다는 선행연구(Starke et al., 2022; Lünich & Kieslich, 2024)의 결과와도 맥락을 같이한다.

셋째, 피드백 제공의 효과가 평가자 유형에 따라 상이하게 작용함을 확인하였다. 특히 AI 평가 상황에서는 피드백 제공이 평가 과정에 대한 이해를 높이고 불확실성을 완화함

으로써 공정성 인식을 향상시키는 중요한 요인으로 작용할 수 있는 것으로 나타났다. 이는 AI 평가가 지닌 상대적인 불투명성과 기계적 판단에 대한 우려를 피드백이 보완해주는 역할을 함을 시사한다(Starke et al., 2022). 반면 교수자 평가에서는 피드백 제공 여부가 공정성 인식에 큰 영향을 미치지 않거나, 오히려 부정적인 인식을 유발할 가능성도 나타났다. 이는 교수자 평가자에 대한 기본적인 신뢰가 이미 형성되어 있기 때문에 피드백의 추가적 효과가 제한적일 수 있으며, 피드백의 내용이나 방식에 따라 평가의 공정성에 대한 재해석이 이루어질 가능성으로 설명될 수 있다(김소연 외, 2021; 이경자, 유정민, 2021). 따라서 이러한 결과는 AI 기반 평가에서는 공정성 인식을 제고하기 위한 핵심 요소로서 피드백 제공을 적극적으로 고려할 필요가 있으며, 교수자 평가에서는 피드백의 제공 여부보다는 그 질과 전달 방식에 대한 보다 정교한 접근이 요구됨을 시사한다.

이러한 결과를 바탕으로 볼 때, 본 연구는 AI 기반 평가의 수용성을 논의함에 있어 기술적 성능이나 채점 정확성에 대한 논의를 넘어, 학습자가 인식하는 공정성이라는 심리적·인지적 요소를 통합적으로 고려할 필요가 있음을 제시한다는 점에서 의의를 갖는다. 특히 평가 공정성은 평가자 유형 자체보다도 학습자의 AI에 대한 신뢰와 피드백과 같은 설명적 정보 제공 여부에 의해 유동적으로 형성된다는 점에서, AI 평가 시스템 설계 시 단순한 자동채점 기능을 넘어 설명가능성과 피드백 중심 설계를 핵심 요소로 고려할 필요성이 제안된다. 또한 본 연구는 동일한 평가 결과라도 평가 맥락과 학습자의 인식에 따라 공정성 판단이 달라질 수 있음을 실증적으로 보여줌으로써, 향후 AI 평가 도입 시 학습자의 신뢰 형성과 평가 과정의 투명성 확보가 필수적임을 강조하였다. 나아가 교육현장에서 AI와 교수자 평가를 이분법적으로 대체 관계로 보기보다는, AI의 객관성과 교수자의 맥락적 판단 및 피드백을 상호 보완적으로 결합하는 하이브리드 평가 체제의 필요성을 제안한다는 점에서도 중요한 시사점을 제공한다.

본 연구는 AI 평가에 대한 다음과 같은 교육적·실천적 시사점을 제공한다. 첫째, 교육현장에서 AI 평가를 도입할 때에는 평가목적에 따른 차이가 일부 존재하더라도, 평가기준의 일관성, 설명 가능한 피드백, 평가 과정과 기준에 대한 명확한 설명이 학습자의 공정성 인식을 높이는 핵심 요인임을 보여준다. 둘째, 학습자의 AI 신뢰 수준을 고려하지 않은 평가는 오히려 공정성 인식에 부정적인 영향을 미칠 수 있으므로, 학교 차원에서 AI 리터러시 함양과 신뢰 형성을 지원하기 위해 AI 기본 소양교육과 AI 활용 탐구 역량 강화, 문제 해결 중심의 교육 프로그램을 체계적으로 개발하고 지원할 필요가 있다. 셋째, AI 평가는 교수자 평가를 대체하기보다 보완적 도구로 활용될 필요가 있다. 특히 객관적 기준이 적용되는 영역과 맥락적 해석이 필요한 영역을 구분하여 각각에 적합한 평가 주체를 배치함으로써, 평가의 신뢰성과 타당성을 동시에 확보할 수 있는 평가 체제를 제안할 수 있다. 예를 들어, 1차 단계에서는 AI가 평가 기준에 따라 일관된 채점

을 수행하고, 2차 단계에서는 교수자가 학습자의 수행 맥락과 피드백을 검토·보정하며, 최종 단계에서는 두 평가 결과를 종합적으로 판단하는 체제라고 할 수 있다.

마지막으로, 본 연구의 한계와 후속 연구를 제안하면 다음과 같다. 첫 째, 본 연구에서는 평가목적에 대한 조작점검을 별도로 실시하지 않아, 참여자들이 진단·형성·총괄 평가 간 차이를 충분히 인식했는지를 직접적으로 확인하지 못하였다. 이에 따라 평가목적에 따른 공정성 인식 차이가 유의하게 나타나지 않은 결과는 실제 차이가 미미했을 가능성과 함께, 시나리오 간 구분이 충분히 이루어지지 않았을 가능성을 함께 고려할 필요가 있다. 향후 연구에서는 조작점검 문항을 포함하고 평가목적 간 차이를 보다 명확히 드러내는 시나리오 설계를 적용할 필요가 있을 것이다. 둘 째, 본 연구는 편의표집을 사용하여 사범계열 학생의 비율이 높아 표본의 대표성이 제한될 수 있다. 특히 이들의 평가에 대한 높은 관심과 이해가 AI 평가에 대한 공정성 인식에 영향을 미쳤을 가능성이 있으므로, 결과를 전체 대학생이나 초·중등 교육 맥락으로 일반화하는 데에는 주의가 필요하다. 셋 째, 본 연구의 시나리오는 평가의 목적과 맥락이 명확히 구분되게 구성하였고 또한 실제 수업 맥락에서 경험되는 평가 흐름(진단→형성→총괄)을 반영하도록 설계함으로써 응답자들에게 동일한 순서로 시나리오를 제시하였다. 그러나 응답이 선행 시나리오의 영향을 받았을 가능성을 완전히 배제하기 어렵다는 점에서 향후 연구에서는 시나리오 제시 순서를 무선화하거나 역균형화 설계를 적용하여 순서효과를 보다 엄밀하게 통제할 필요가 있다. 넷 째, 상호작용 효과의 시각화를 위해 AI 신뢰를 ±1SD 기준으로 구분하여 제시하였으며, 이러한 방식은 연속변인의 정보 손실과 검정력 저하를 초래할 가능성이 있다(MacCallum et al., 2002). 향후 연구에서는 Johnson-Neyman 기법 등 연속변인 기반 사후분석을 적용하여 상호작용 효과를 보다 정교하게 파악할 필요가 있다. 다섯 째, 본 연구는 세 조절변인을 독립적으로 분석하였으므로, 통합 모형에서의 효과 추정치 변화 가능성은 후속 연구에서 검토할 필요가 있다. 한편, 본 연구의 제한점을 확장하여 향후 연구에서는 실제 AI 평가 시스템이 적용된 교실 맥락에서의 실험연구를 수행하고, 다양한 연령 및 학습자를 포함한 연구를 통해 공정성 인식의 맥락적 차이를 탐색할 필요가 있다. 더 나아가, AI 평가에 대한 공정성 인식이 기술수용태도, 평가태도, 학습방식에 미치는 영향을 통합적으로 규명함으로써 디지털 교실 환경에서 신뢰 가능한 평가체제 구축에 기여할 수 있을 것이다.

참고문헌

경기도교육청 (2024). **경기도교육청 AI 기반 교수·학습 플랫폼 효과성 분석 및 개선 방안 연구보고서**. 경기도교육연구원.

고경민, 이용상, 신동광, 이효신 (2025). 교양 글쓰기 교육 강화를 위한 자동평가 적용 사례: K 대학 사례를 중심으로. **교육문화연구, 31**(4), 605-619.

교육부 (2022). 디지털 인재 양성 종합방안. 교육부.

교육부 (2023). 디지털 기반 교육혁신 방안. 교육부.

교육부 (2024). 2024년 주요 정책 추진계획. 교육부.

교육부 (2025). 2025년 주요 업무 추진계획. 교육부.

권은아, 이종구 (2020). 상황판단검사 형식을 이용한 정직성검사의 타당화. **한국심리학회지: 산업 및 조직, 33**(4), 545-569.

김소연, 김창희, 박한승, 김채원, 오윤정, 최진석 (2021). COVID-19 이후 대학 E-learning 만족도 요인 분석 : IPA 분석을 중심으로. **글로벌경영학회지, 18**(3), 133-154.

김은영 (2024). 일본어 작문 수업에서 챗 GPT 활용에 대한 학습자의 인식 및 태도 연구. **일본어교육, 107**, 37-49.

김재욱, 손원숙 (2021). 형성적 피드백 과정에서 피드백 환경의 역할 탐색. **교육평가연구, 34**(1), 31-52.

김주현 (2025). 정보 텍스트 기반 초등영어 수업에서의 효과적인 하이러닝 활용 방안 . 석사학위논문, 서울교육대학교.

박도순, 홍후조 (2010). **교육과정과 교육평가.** 문음사.

박소영, 이병윤, 함은혜, 이유경, 이성혜 (2023). ChatGPT-4 의 과학적 탐구 역량 평가 가능성 탐색: 인간평가자와의 비교를 중심으로. **교육학연구, 61**(4), 299-332.

백순근, 박현정, 이다경, 양현경, 김서진 (2024). 시나리오 기반 척도의 응답 편향성 개선 효과 분석: 리커트형 척도와의 비교를 중심으로. **아시아교육연구, 25**(3), 343-373.

안혜정 (2024). 학교현장실습학기제에 참여한 예비과학교사들이 경험하는 수업에서 나타나는 교실대화 분석. **교육과정평가연구, 27**(2), 231-250.

오의경 (2025). 대학생의 AI 활용 특성과 인식 분석. **한국문헌정보학회지, 59**(1), 671-692.

윤가율 (2024). 중등 예비교사의 생성형 AI 활용 평가에 대한 수용 및 영향요인 간의 구조

적 관계 분석. 석사학위논문, 경북대학교.

이경자, 유정민 (2021). 대학생들의 비대면 수업 평가 방식 공정성에 대한 인식 연구 : I 대학의 비대면 수업 평가 공정성에 대한 인식 진단을 중심으로. **교양교육연구, 15**(6), 301-318.

이빛나, 손원숙 (2018). 피드백 효과에 대한 메타분석. **교육평가연구, 31**(3), 501-529.

이빛나, 박민애 (2023). 국어 수업에서 고등학생의 형성평가 인식 유형 및 영향 요인 탐색. **교육과정평가연구, 26**(4), 93-111.

정유남 (2024). 대학 글쓰기 수업에서 학생 인식 기반 AI 피드백의 효용성 연구. **교양교육연구, 18**(5), 159-173.

정윤하 (2025). 서·논술형 평가를 위한 생성형 AI 기반 자동채점 프로그램 개발 및 타당화. 박사학위논문, 서울대학교.

하민수, 신철균 (2024). 교양교육을 위한 인공지능 활용 서· 논술형 평가 및 피드백 적용 사례 · 지속가능발전의 이해 기초교양 강좌를 중심으로. **교양교육연구, 18**(2), 11-22.

하유라 (2021). 대학생의 피드백 추구 행동과 학업성취와의 관계:피드백 추구 동기 및 자기성찰의 역할. **교육과정평가연구, 24**(2), 127-145.

함은혜, 박소영, 이병윤, 이성혜, 이유경, 홍유정 (2024). GPT-4 를 활용한 과학탐구역량 자동채점의 특성 분석. **교육정보미디어연구, 30**(3), 713-742.

홍소영 (2018). 학생 자기평가의 학습효과에 관한 메타분석. **교육평가연구, 31**(1), 309-331.

Aguinis, H., & Bradley, K. J. (2014). Best practice recommendations for designing and implementing experimental vignette methodology studies. *Organizational research methods, 17*(4), 351-371.

Atzmüller, C., & Steiner, P. M. (2010). Experimental vignette studies in survey research. *Methodology. 6*(3), 128-138

Bloom, B. S. (1969). Some theoretical issues relating to educational evaluation. *Teach. Coll. Rec.* 70, 26-50.

Bloom, B. S., Hastings, J. T., & Madaus, G. F. (1971). *Handbook on formative and summative evaluation of student learning*. New York, NY: McGraw-Hill.

Bui, N. M., & Barrot, J. S. (2025). ChatGPT as an automated essay scoring tool in the writing classrooms: how it compares with human scoring. *Education and*

Information Technologies, 30(2), 2041-2058.

Colquitt, J. A. (2001). On the dimensionality of organizational justice: *a construct validation of a measure. Journal of Applied Psychology, 86*(3), 386-400.

Chai, F., Ma, J., Wang, Y., Zhu, J., & Han, T. (2024). Grading by AI makes me feel fairer? How different evaluators affect college students' perception of fairness. *Frontiers in Psychology, 15*, 1221177.

Chinta, S. V., Wang, Z., Yin, Z., Hoang, N., Gonzalez, M., Quy, T. L., & Zhang, W. (2024). FairAIED: Navigating fairness, bias, and ethics in educational AI applications. arXiv preprint arXiv: 2407.18745.

Dahri, N. A., Yahaya, N., Al-Rahmi, W. M., Aldraiweesh, A., Alturki, U., Almutairy, S., Shutaleva, A., & Soomro, R. B. (2024). Extended TAM based acceptance of AI-Powered ChatGPT for supporting metacognitive self-regulated learning in education: A mixed-methods study. *Heliyon, 10*(8), e29317

Ekici, H., Yücel, E., & Cesur, S. (2023). Deciding between moral priorities and COVID-19 avoiding behaviors: A moral foundations vignette study. *Current Psychology, 42*(7), 5922-5938.

Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Shum, S. B., & Koedinger, K. R. (2022). Ethics of AI in education: Towards a community-wide framework. *International Journal of Artificial Intelligence in Education, 32*(3), 504-526.

Liu, L. T., Wang, S., Britton, T., & Abebe, R. (2023). Reimagining the machine learning life cycle to improve educational outcomes of students. *Proceedings of the National Academy of Sciences, 120*(9), e2204781120.

Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). Intelligence unleashed: An argument for AI in education. London: Pearson.

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7*(1), 19.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional science, 18*(2), 119-144.

Shin, D. (2021). The effects of explainability and causability on perception, trust, and

acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies*,146, 102551.

Starke, C., Baleis, J., Keller, B., & Marcinkowski, F. (2022). Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. *Big Data & Society, 9*(2), 20539517221115189.

Tierney, R. D. (2012). Fairness in classroom assessment. SAGE handbook of research on classroom assessment(pp. 125-144). Thousand Oaks, CA, United States; SAGE Publications.

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view1. *MIS Quarterly, 27*(3), 425-478.

Winne, P. H., & Butler, D. L. (1994). Student cognition in learning from teaching. International Encyclopedia of Education, 2, 5738-5775.

논문접수 : 2026.4.2. / 수정본 접수 : 2026.4.28. / 게재승인 : 2026.5.11.

ABSTRACT

College Students' Perceptions of Assessment Fairness in AI-Based and Instructor-Led Evaluation: Moderating Roles of Assessment Purpose, Trust in AI, and Feedback

Haesook Kim
Ph.D Candidate, Kyungpook National University
Wonsook Sohn
Professor, Kyungpook National University

As AI-based evaluation systems are increasingly adopted in higher education, concerns remain about whether students perceive AI scoring as fair compared with instructor-led evaluation. This study examined college students' perceptions of assessment fairness in AI-based versus instructor-led evaluation in a general education writing course and whether this relationship varied by assessment purpose (diagnostic, formative, summative), trust in AI, and feedback provision. Three independent survey studies were conducted with undergraduates from four-year universities in South Korea (Study 1: N = 390; Study 2: N = 267; Study 3: N = 253) using convenience sampling in which participants provided repeated fairness ratings for vignettes. To account for within-person dependence across repeated judgments and between-person variability, two-level mixed-effects models with participants as random intercepts were fitted. The results indicated that, first, college students perceived AI-based evaluation as significantly fairer than instructor-led evaluation; however, this association did not vary by assessment purpose. Second, trust in AI significantly moderated the relationship between evaluator type and perceived fairness, such that higher trust in AI was associated with significantly higher perceived fairness of AI evaluation. Third, feedback provision also significantly moderated this relationship, with perceived fairness of AI evaluation being higher when feedback was provided than when it was not. Finally, these findings suggest that designing and implementing AI-based assessment should take into account diverse psychological and contextual factors related to the formation of

fairness perceptions, and educational and practical implications are discussed.

Key Words: AI-based Evaluation, Instructor-led Evaluation, Assessment Purpose, Assessment Fairness, Trust in AI, Feedback